

Claims

1. A noise reduction system with an audio-visual user interface, said system being specially adapted for running an application for combining visual features ( $\varrho_{v,nT}$ ) extracted from a digital video sequence ( $v(nT)$ ) showing the face of a speaker ( $S_i$ ) with audio features ( $\varrho_{a,nT}$ ) extracted from an analog audio sequence ( $s(t)$ ), wherein said audio sequence ( $s(t)$ ) can include noise in the environment of said speaker ( $S_i$ ), said noise reduction system (200b/c) comprising

- means (101a, 106b) for detecting and analyzing said analog audio sequence ( $s(t)$ ),
- means (101b') for detecting said video sequence ( $v(nT)$ ), and
- means (104a+b, 104'+104'') for analyzing the detected video signal ( $v(nT)$ ), characterized by

a noise reduction circuit (106) being adapted to separate the speaker's voice from said background noise ( $n'(t)$ ) based on a combination of derived speech characteristics ( $\varrho_{av,nT}$

15 :=  $[\varrho_{a,nT}^T, \varrho_{v,nT}^T]^T$ ) and outputting a speech activity indication signal ( $\hat{s}_i(nT)$ ) which is obtained by a combination of speech activity estimates supplied by said analyzing means (106b, 104a+b, 104'+104'').

2. A noise reduction system according to claim 1,

20 characterized by

means (SW) for switching off an audio channel in case the actual level of said speech activity indication signal ( $\hat{s}_i(nT)$ ) falls below a predefined threshold value.

3. A noise reduction system according to anyone of the claims 1 or 2,

25 characterized by

a multi-channel acoustic echo cancellation unit (108) being specially adapted to perform a near-end speaker detection and double-talk detection algorithm based on acoustic-phonetic speech characteristics derived by said audio feature extraction and analyzing means (106b) and said visual feature extraction and analyzing means (104a+b, 104'+104'').

4. A noise reduction system according to anyone of the claims 1 to 3,  
characterized in that  
said audio feature extraction and analyzing means (106b) is an amplitude detector.

5 5. A near-end speaker detection method reducing the noise level of a detected analog audio sequence ( $s(t)$ ),

said method being characterized by the following steps:

- subjecting (S1) said analog audio sequence ( $s(t)$ ) to an analog-to-digital conversion,
- calculating (S2) the corresponding discrete signal spectrum ( $S(k \cdot \Delta f)$ ) of the analog-to-digital-converted audio sequence ( $s(nT)$ ) by performing a Fast Fourier Transform (FFT),
- detecting (S3) the voice of said speaker ( $S_i$ ) from said signal spectrum ( $S(k \cdot \Delta f)$ ) by analyzing visual features ( $\underline{\varrho}_{v,nT}$ ) extracted from a simultaneously with the recording of the analog audio sequence ( $s(t)$ ) recorded video sequence ( $v(nT)$ ) tracking the current location of the speaker's face, lip movements and/or facial expressions of the speaker ( $S_i$ ) in subsequent images,
- estimating (S4) the noise power density spectrum ( $\Phi_{nn}(f)$ ) of the statistically distributed background noise ( $\tilde{n}(t)$ ) based on the result of the speaker detection step (S3),
- subtracting (S5) a discretized version ( $\tilde{\Phi}_{nn}(k \cdot \Delta f)$ ) of the estimated noise power density spectrum ( $\tilde{\Phi}_{nn}(f)$ ) from the discrete signal spectrum ( $S(k \cdot \Delta f)$ ) of the analog-to-digital-converted audio sequence ( $s(nT)$ ), and
- calculating (S6) the corresponding discrete time-domain signal ( $\hat{s}_i(nT)$ ) of the obtained difference signal by performing an Inverse Fast Fourier Transform (IFFT), thereby yielding a discrete version of the recognized speech signal.

25 6. A near-end speaker detection method according to claim 5,  
characterized by the step of

conducting (S7) a multi-channel acoustic echo cancellation algorithm which models echo path impulse responses by means of adaptive finite impulse response (FIR) filters and subtracts echo signals from the analog audio sequence ( $s(t)$ ) based on acoustic-phonetic speech

30 characteristics derived by an algorithm for extracting visual features ( $\underline{\varrho}_{v,nT}$ ) from a video

sequence ( $v(nT)$ ) tracking the location of a speaker's face, lip movements and/or facial expressions of the speaker ( $S_i$ ) in subsequent images.

7. A near-end speaker detection method according to claim 6,

5 characterized in that

said multi-channel acoustic echo cancellation algorithm performs a double-talk detection procedure.

8. A near-end speaker detection method according to anyone of the claims 5 to 7,

10 characterized in that

said acoustic-phonetic speech characteristics are based on the opening of a speaker's mouth as an estimate of the acoustic energy of articulated vowels or diphthongs, respectively, rapid movement of the speaker's lips as a hint to labial or labio-dental consonants, respectively, and other statistically detected phonetic characteristics of an association between

15 position and movement of the lips and the voice and pronunciation of said speaker ( $S_i$ ).

9. A near-end speaker detection method according to anyone of the claims 5 to 8,

characterized by

a learning procedure used for enhancing the step of detecting (S3) the voice of said speaker ( $S_i$ ) from the discrete signal spectrum ( $S(k\Delta f)$ ) of the analog-to-digital-converted version ( $s(nT)$ ) of an analog audio sequence ( $s(t)$ ) by analyzing visual features ( $\varrho_{v,nT}$ ) extracted from a simultaneously with the recording of the analog audio sequence ( $s(t)$ ) recorded video sequence ( $v(nT)$ ) tracking the current location of the speaker's face, lip movements and/or facial expressions of the speaker ( $S_i$ ) in subsequent images.

25

10. A near-end speaker detection method according to anyone of the claims 5 to 9,

characterized by the step of

correlating (S8a) the discrete signal spectrum ( $S_\tau(k\Delta f)$ ) of a delayed version ( $s(nT-\tau)$ ) of the analog-to-digital-converted audio signal ( $s(nT)$ ) with an audio speech activity estimate obtained by an amplitude detection (S8b) of the band-pass-filtered discrete signal spectrum ( $S(k\Delta f)$ ), thereby yielding an estimate ( $\tilde{S}_i(f)$ ) for the frequency spectrum ( $S_i(f)$ ) corre-

30

sponding to the signal ( $s_i(t)$ ) which represents said speaker's voice as well as an estimate ( $\tilde{\Phi}_{nn}(f)$ ) for the noise power density spectrum ( $\Phi_{nn}(f)$ ) of the statistically distributed background noise ( $n'(t)$ ).

- 5 11. A near-end speaker detection method according to claim 10,  
characterized by the step of

correlating (S9) the discrete signal spectrum ( $S_i(k\cdot\Delta f)$ ) of a delayed version ( $s(nT-\tau)$ ) of the analog-to-digital-converted audio signal ( $s(nT)$ ) with a visual speech activity estimate taken from a visual feature vector ( $\varrho_{v,i}$ ) supplied by the visual feature extraction and analyzing means (104a+b, 104'+104''), thereby yielding a further estimate ( $\tilde{S}_i(f)$ ) for updating the estimate ( $\tilde{S}_i(f)$ ) for the frequency spectrum ( $S_i(f)$ ) corresponding to the signal ( $s_i(t)$ ) which represents said speaker's voice as well as a further estimate ( $\tilde{\Phi}_{nn}'(f)$ ) for updating the estimate ( $\tilde{\Phi}_{nn}(f)$ ) for the noise power density spectrum ( $\Phi_{nn}(f)$ ) of the statistically distributed background noise ( $n'(t)$ ).

15

12. A near-end speaker detection method according anyone of the claims 10 or 11,  
characterized by the step of

adjusting (S10) the cut-off frequencies of a band-pass filter (204) used for filtering the discrete signal spectrum ( $S(k\cdot\Delta f)$ ) of the analog-to-digital-converted audio signal ( $s(t)$ ) de-

- 20 pendent on the bandwidth of the estimated speech signal spectrum ( $\tilde{S}_i(f)$ ).

13. A near-end speaker detection method according to anyone of the claims 5 to 9,  
characterized by the steps of

– adding (S11a) an audio speech activity estimate obtained by an amplitude detection of  
25 the band-pass-filtered discrete signal spectrum ( $S(k\cdot\Delta f)$ ) of the analog-to-digital-converted audio signal ( $s(t)$ ) to a visual speech activity estimate taken from a visual feature vector ( $\varrho_{v,i}$ ) supplied by said visual feature extraction and analyzing means (104a+b, 104'+104''), thereby yielding an audio-visual speech activity estimate,

– correlating (S11b) the discrete signal spectrum ( $S(k\cdot\Delta f)$ ) with the audio-visual speech activity estimate, thereby yielding an estimate ( $\tilde{S}_i(f)$ ) for the frequency spectrum ( $S_i(f)$ )

corresponding to the signal ( $s_i(t)$ ) which represents said speaker's voice as well as an estimate ( $\tilde{\Phi}_{nn}(f)$ ) for the noise power density spectrum ( $\Phi_{nn}(f)$ ) of the statistically distributed background noise ( $n'(t)$ ) and

- adjusting (S11c) the cut-off frequencies of a band-pass filter (204) used for filtering the discrete signal spectrum ( $S(k\Delta f)$ ) of the analog-to-digital-converted audio signal ( $s(t)$ ) dependent on the bandwidth of the estimated speech signal spectrum ( $\tilde{S}_i(f)$ ).

14. Use of a noise reduction system (200b/c) according to anyone of the claims 1 to 4 and a near-end speaker detection method according to anyone of the claims 5 to 13 for a video-  
10 telephony based application in a telecommunication system running on a video-enabled phone with a built-in video camera (101b') pointing at the face of a speaker ( $S_i$ ) participating in a video telephony session.

15. A telecommunication device equipped with an audio-visual user interface,  
15 characterized by  
noise reduction system (200b/c) according to anyone of the claims 1 to 4.